Accidental Deaths due to Drug Overdose Analytics

Mayank Grover
College of Engineering and Computing
George Mason University
Fairfax, Virginia
mgrover3@gmu.edu

***Abstract*— Drug Overdose is a hot topic of concern for various governments across the world. The world battles to reduce the deaths occurring due to drug overdose. In this work, I propose to present the statistics and number of deaths occurring in the state of Connecticut, United States of America. I will be using certain features of Amazon Web Services (AWS), Python and R to perform the exploratory data analysis of the Accidental Death due to Drug Overdose dataset [1].**

***Keywords*— Opioid, Drugs, Drug Overdose, Deaths, Analytics**

1. INTRODUCTION-

With the increase in the rates of death across the globe due to overdose of external substances in human blood, we must pay attention to the cause of the concern on an immediate basis. Various substances are out there in the market which are consumed by people in large amount which lead to fatal health conditions. There are known substances like heroin, cocaine, etc., and unknown substances which are made up of different compositions of chemicals. Not just harmful drugs, there are general medications as well which people overdose on as a part of self-treating some illnesses. All these causes together raise a concern for the medical agencies and the government agencies.

Increases in Drug and Opioid Overdose Deaths—United States, 2000–2014 [2] compares the change in death rate from the year 2000 to 2014 across all the states in the US. It uses demographic data such as age, sex, race Hispanic origin, and the states of origin. It compares the death numbers due to different types of drugs such as heroin, methadone, synthetic opioids excluding methadone, natural and semi-synthetic opioids, and drugs involving opioids. This paper presents the statistics of drug overdose and the accidental deaths that occurred in the state of Connecticut, USA. As a result of this report, I want to extract answers to the following research questions: 'What type of drug causes what kind of death, and the count of deaths each drug caused to date?', 'What is the city-wise average age of people dying due to drug overdose ', and 'What is the yearly trend of deaths due to drug overdose in different races?'.

Further, this paper is organized into different sections. Section 2 contains a discussion of previously analyzed published articles on the same issue. Section 3 describes the methods used to pre-process the data, analyze, and obtain the results. Section 4 presents the results obtained. Section 5 describes about the conclusion and the future scope of the project work.

2. RELATED WORK-

Opioid Overdose–Related Emergency Department Visits and Accidental Deaths during the COVID-19 Pandemic [3] provides reports about the opioid epidemic throughout recent years, starting in 2016 in San Francisco, from January to April. It provides insights into the number who visited emergency department visits and accidental deaths during the COVID-19 pandemic. The report displays the demographics of the population, which age group they belonged to and what was their gender. The data show the number of deaths and emergency visits in case of opioid overdose.

Accidental drug-induced deaths due to opioids in Australia, 2013 [4] covers the analysis of the death rates among different age groups caused due to overdose of opioids considering both cases, inclusion and non-inclusion of heroin. The period covered for analysis is the duration of 2007 - 2013.

3. METHODS-

This proposed system uses MySQL for Database management, python and R for the analytics. The accidental deaths due to drug overdose dataset contains various columns (Fig 3.1).

```
Data columns (total 49 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Date                              6871 non-null   object
 1   Date.Type                         6871 non-null   object
 2   Age                               6871 non-null   int64
 3   Sex                               6871 non-null   object
 4   Race                              6871 non-null   object
 5   Ethnicity                         1038 non-null   object
 6   Residence.City                    6548 non-null   object
 7   Residence.County                  6473 non-null   object
 8   Residence.State                   6529 non-null   object
 9   Injury.City                       6781 non-null   object
 10  Injury.County                     5905 non-null   object
 11  Injury.State                      4974 non-null   object
 12  Injury.Place                      6814 non-null   object
 13  Description.of.Injury             6126 non-null   object
 14  Death.City                        6870 non-null   object
 15  Death.County                      6839 non-null   object
 16  Death.State                       6871 non-null   object
 17  Location                          6864 non-null   object
 18  Location.if.Other                 500 non-null    object
 19  Cause.of.Death                    6871 non-null   object
 20  Manner.of.Death                   6868 non-null   object
 21  Other.Significant.Conditions      728 non-null    object
 22  Heroin                            2095 non-null   object
 23  Heroin.death.certificate..DC.     399 non-null    object
 24  Cocaine                           2514 non-null   object
 25  Fentanyl                          5055 non-null   object
 26  Fentanyl.Analogue                 728 non-null    object
 27  Oxycodone                         522 non-null    object
 28  Oxymorphone                       66 non-null     object
 29  Ethanol                           1941 non-null   object
```

Fig. 3.1 – Variables in the dataset

Along with the above-mentioned technological stack, I also used the AWS cloud platform to store the data and explore the data using the AWS Glue (Fig 3.2). The null values from this dataset from the 'Sex' column were also removed.

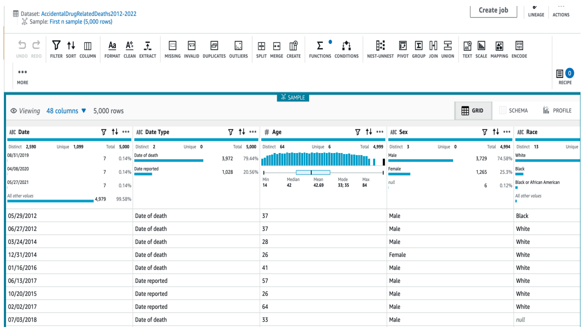

Fig 3.2 – Exploring data using AWS Glue

MySQL was used to filter the data and study the resulting rows. I filtered all the rows where 'Death City' was Connecticut (Fig 3.3).

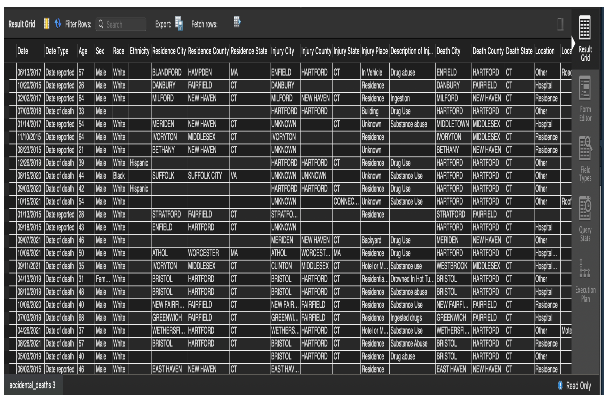

Fig. 3.3 – MySQL Data Exploration

I used the R language to read the dataset from the AWS S3 bucket and perform the univariate analysis of the significant variables. Fig. 3.4 shows age variable was explored in order to study the number of deaths occurring in different age groups throughout the period of 2012-2021 in Connecticut. Fig. 3.5 studies the gender distribution of the population with the death numbers and then with the Race of the population as well.

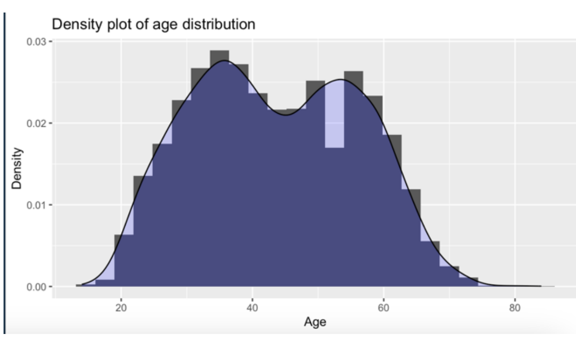

Fig. 3.4 – Density plot to study the death rates at different ages

Fig. 3.6 shows grouping of the data according to the death city was done. Fig 3.7 depicts the plot of year wise deaths recorded in Connecticut.

In the next part of analysis, I used python to generate the plots to study the average age of population in the death city recorded.

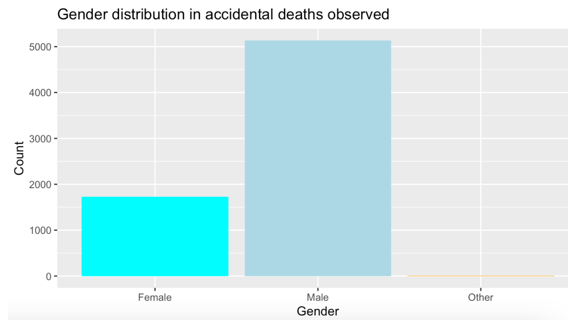

Fig. 3.5 – Gender distribution in accidental deaths due to drug overdose

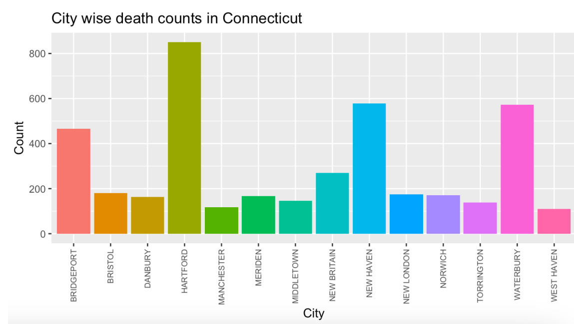

Fig. 3.6 – City wise death count in Connecticut

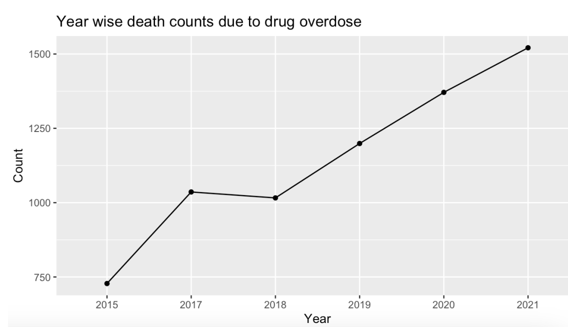

Fig. 3.7 – Year wise death counts due to drug overdose in Connecticut

Line charts were also plotted to study the number of yearly deaths recorded in each race. After the analysis part, I planned to create a model which predicts the substance used based on the four parameters: age, residence city, race and gender. Feature engineering was applied to convert the categorical character variables to numeric data. A heatmap (Fig. 3.6) to study the correlation of variables was plotted and was observed that there is very little correlation between the variables. This helped me to conclude that linear relationship does not exists, but in order to create a model. Random forest approach was used to build a model.

4. RESULTS

In order to study the description of injury caused by each drug, bar charts were plotted to count the frequency of injuries for specific substances. Fig 4.1,4.2 and 4.3 illustrates the top description of injuries caused by methadone, oxycodone, and gabapentin respectively.

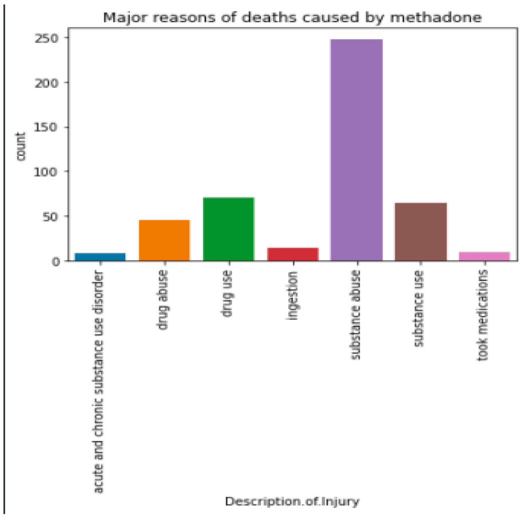

Fig. 4.1- Description of injury caused by methadone

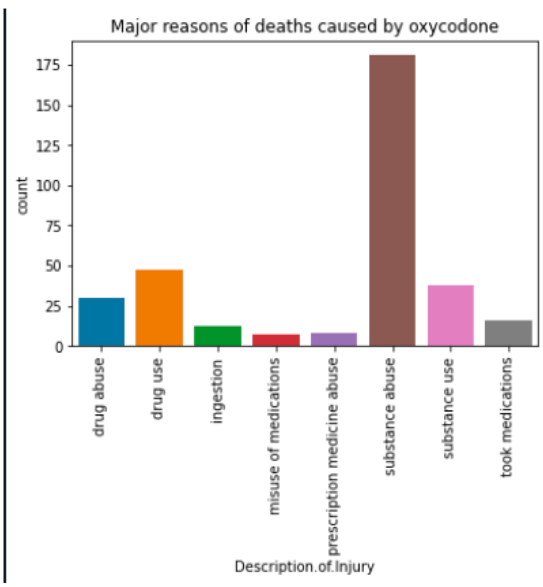

Fig. 4.2- Description of injury caused by oxycodone

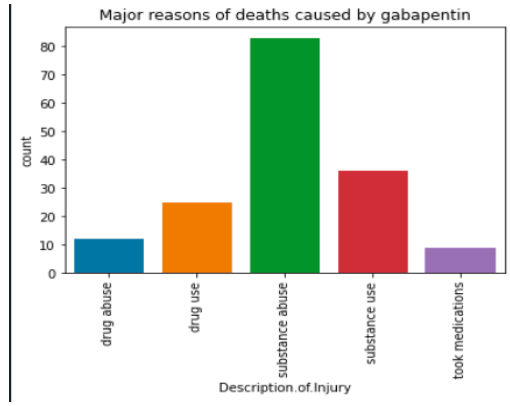

Fig. 4.3- Description of injury caused by gabapentin

The above figures show that the 'substance abuse' is the most common cause of deaths caused by any drug. It is also observed that medication of the drug also caused death which comes as a surprise, since the medications are prescribed by the doctors. Drug abuse and drug use also are one of the common causes of injury caused due to drug overdose.

Fig. 4.4 depicts the year wise death counts in different races.

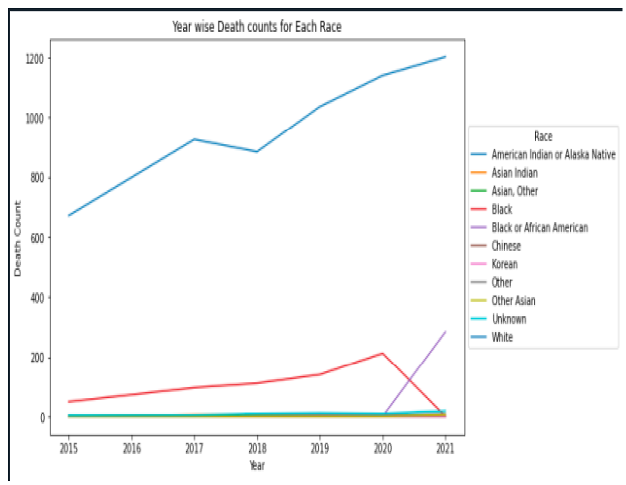

Fig. 4.4 – Yearly death count in various races

From the above chart, it is observed that the greatest increase in deaths was observed in race 'American Indian or Alaska Native'. There was a significant decline in the 'black' race in the number of deaths in the year 2021 due to drug overdose. During the same time, there was a sudden increase in deaths in the 'Chinese' race in the year 2021.

Fig. 4.5 displays all the cities having average age of people dying due to drug overdose accidentally in between 20-40.

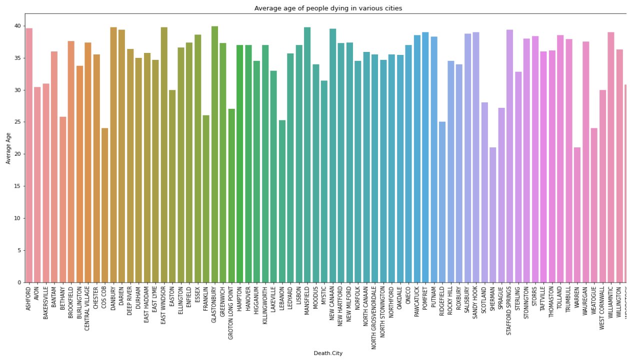

Fig. 4.5- Average age of people dying due to drug overdose accidentally

This provides us with the information that more than 70 cities have average age in between 20-40 to which the people belong who face death due to drug overdose in Connecticut.

Describing about the modelling part, residence city, race, gender and the substance used was converted to numerical data using the categorical assigning of codes. After this step the data was completely numeric, but Fig. 4.6 shows that these variables are not in linearly related to each other, hence could not apply linear regression. I have kept substance as the predicted value and the rest as the predictors. Started with the random forest approach, a model with mean squared error as 28.24 was observed. Fig. 4.6 displays the important variables from the random forest model built.

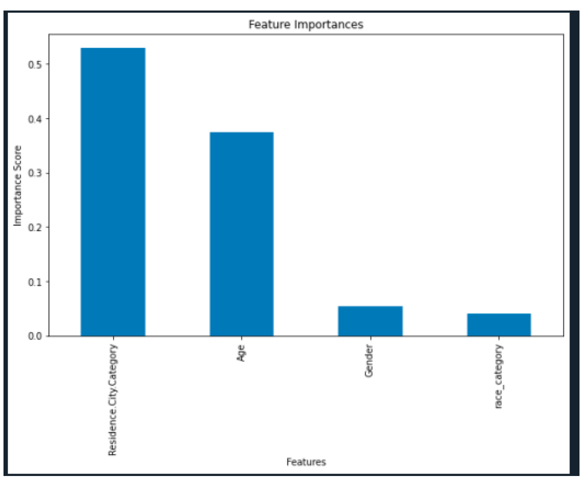

Fig. 4.6 Important variables from the random forest model

5. CONCLUSIONS

This report provides insights and analysis of the accidental deaths occurring due to drug overdose in the state of Connecticut, USA. Complete demographic results were displayed for the reader to understand the causes of drug overdose, the average age group which becomes a victim of drug abuse or drug use. I also created a model in order to predict the substance used based on the inputs provided.

In future works, this research can be extended by analyzing more specific variables such as ethnicity, injury place and location of death. Modelling can be improved to get better results using other machine learning techniques such as gradient boosting to build a better model with higher accuracy.

6. REFERENCES

[1] Pauline Zaldonis, Accidental Drug Related Deaths 2012-2022, data.ct.gov[publisher], 2022-05-13. https://catalog.data.gov/dataset/accidental-drug-related-deaths-2012- 2018

[2] R.A. Rudd and N. Aleshire and J.E. Zibbell and R. {Matthew Gladden}. Increases in Drug and Opioid Overdose Deaths—United States, 2000–2014. Journal - American Journal of Transplantation. Available: https://doi.org/10.1111/ajt.13776

[3] Rodda, L.N., West, K.L. & LeSaint, K.T. Opioid Overdose–Related Emergency Department Visits and Accidental Deaths during the COVID-19 Pandemic. J Urban Health 97 , 808–813 (2020).Available: https://doi.org/10.1007/s11524-020-00486-y

[4] apo.org.au, 'Accidental drug-induced deaths due to opioids in Australia, 2013' - Analysis & Policy Observatory (APO), licensed under a Creative Commons Attribution-Non-Commercial 3.0 Australia (CC-BY-NC 3.0 AU) License. Available online: https://apo.org.au/node/99646